

# **METHOD AND DEVICE FOR CLASSIFYING INTERNET OBJECTS AND OBJECTS STORED ON COMPUTER-READABLE MEDIA**

## **FIELD OF THE INVENTION**

This invention relates to a method and device for classifying Internet objects, and/or objects stored on computer-readable media. More particularly, this invention relates to a method and device for classifying Internet objects on the basis of adult content, using an automated web robot, as well as a method and device for classifying objects stored on computer-readable media, using a classification module stored on a computer system.

## **BACKGROUND OF THE INVENTION**

An Internet object is anything that is downloaded or transmitted via the Internet, including but not limited to web pages, images, text documents, email messages, newsgroup postings, chat text, video, and audio. Given the tremendous increase in the size and variety of the Internet community in recent years, classification of Internet objects has become increasingly important, and manual classification of Internet objects is becoming increasingly inadequate.

Internet objects can be classified by the subject matter they contain. One practical application of such classification is the ability to filter Internet objects based on classification. There is particular interest in classifying Internet objects containing adult content because access to such Internet objects can then be restricted to prevent viewing by minors.

Internet filtering products have previously been developed to attempt to filter Internet objects based on adult content. All filtering products require a method by which to classify Internet objects. The prior art methods of classification are detailed below, but can be summarized as taking one of three approaches: (i) filtering based on classification information embedded in an Internet object; (ii) compilation of "blacklists" and/or "whitelists" that filtering products may reference; and (iii) real-time textual analysis.

One Internet filtering product is produced by Netscape as part of the Netscape web browser. The browser includes an adult content filtering feature that classifies web pages based on the PICS labeling system, a voluntary system by which Internet content providers include special codes in their Internet objects (in this case web pages) to describe the content of the object. The PICS labels are the only mechanism used by Netscape in classifying the adult content of web pages. The PICS system is described in greater detail at <http://www.w3.org/pics/>. The drawback in relying solely on the PICS label to classify Internet objects is that not all sites are classified using the system, as participation in the PICS system is voluntary. In addition, there is no independent verification process, and users rely solely on the judgement of the Internet content providers, which may be biased by self-interest.

Cyber Patrol is another Internet filtering product. Cyber Patrol maintains a "blacklist" of web sites that are considered to contain adult content. The Cyber Patrol web page at <http://www.cyberpatrol.com/> discloses that "professional

researchers compile" the lists, apparently manually. With the current growth rate of Internet users and Internet content providers, the current method of manual classification is inadequate.

SurfWatch [<http://www1.surfwatch.com/>] is another Internet filtering product that works by maintaining blacklists. SurfWatch appears to search web pages and URLs for restricted keywords. Any page or URL containing a restricted keyword is classified as adult content. There is no further initial verification process. This can lead to a site being erroneously classified as adult, as illustrated in the recent incident in which one of the pages on the "Whitehouse for Kids" website was classified as adult content because it was named "couples.html".

CYBERsitter is yet another Internet filtering product that attempts to classify web sites, by looking at the text of the page and URLs. The product removes profane English words from the text of web pages, but does not filter out pornographic images from web pages which do not contain text and does not filter out words that are profane words in foreign languages.

NetNanny [<http://www.netnanny.net/>] is still another Internet filtering product that uses a blacklist of domain names of web sites not suitable for children. The NetNanny web site discloses that the NetNanny site list is compiled by NetNanny staff, the suggestions of customers, and third party children's advocacy groups. The ability of the product to filter out undesirable sites is limited by the comprehensiveness of the blacklist, which is compiled manually.

In sum, given the rapid proliferation of Internet objects, manual classification of Internet objects is an inadequate method of classification. Similarly, the use of unweighted or unverified text filtering alone results in inadequate and often inaccurate classification of Internet objects. Given the growing availability of adult content on computer-readable media, there is also need for a method and device that can more accurately and efficiently identify adult content on computer readable media, and either filter or deny access to such adult content.

The present invention can also be used to classify Internet objects and/or objects stored on computer readable media based on other criteria besides adult content.

## SUMMARY OF THE INVENTION

An object of the present invention is to address the limitations of the prior art with respect to classification methods and devices for Internet objects (i.e. anything that is downloaded or transmitted via the Internet, including but not limited to web pages, images, text documents, email messages, newsgroup postings, chat text, video, and audio). The present invention addresses the limitations of the prior art by producing an automated method and device which more accurately classifies Internet objects.

Another object of the present invention is to build a database of Internet URLs (universal resource locators) that contain Internet objects that exceed a

tolerated threshold for adult content, for use upon executing Internet access via any communications port in any type of computer in order to filter or block access to Internet objects that contain adult content.

The above objects are met by the method and device of the invention, which uses weighted textual analysis supplemented by an analysis of Internet content other than text content (such as images, video, link relations, and audio) which are relevant and useful to properly classify an Internet object.

Accordingly, the invention relates to a method and device that classifies an Internet object by calculating a coefficient of adult content ("CoAC") using the following steps:

- (a) computing the following individual data type coefficients (i) the descriptor coefficient, (ii) the name coefficient, (iii) the text coefficient, (iv) the image coefficient, (v) the audio coefficient, (vi) the video coefficient, (vii) the plug-in coefficient, and (viii) the relationship coefficient, and calculating a weighted number to describe the CoAC;
- (b) building and maintaining a database of URLs which contain Internet objects that exceed the tolerated threshold for adult content, for use upon executing Internet access via any communications port in any type of computer for the purpose of filtering access to such identified URLs.

The method can be used to classify Internet objects based on any combination of the above data type coefficients, and possibly other criteria.

Another object of the present invention is to address the limitations of prior art with respect to classification methods and devices for objects stored on

computer-readable media, such as floppy discs, hard discs, CD-ROM, CD-R, CDRW, and any other computer-readable information storage device. In this context, "object" refers to anything that can be stored on such memory devices, including but not limited to images, text, video, and audio objects. The present invention addresses the limitations of the prior art with respect to classification methods and devices for objects stored on computer-readable media by producing an automated method and device which more accurately classifies objects stored on computer-readable media.

Another object of the present invention is to identify objects stored on computer-readable media that exceed a tolerated threshold for adult content, in order to filter or deny computer access to such objects.

The above objects are met by the method and device of the invention, which uses weighted textual analysis supplemented by an analysis of content of the object stored on computer-readable media other than text content (such as images, video, and audio) which are relevant and useful to properly classify an Internet object.

Accordingly, the invention relates to a method and device that classifies an object stored on computer-readable media by calculating a coefficient of adult content ("CoAC") using the following steps:

- (a) computing the following individual data type coefficients (i) the text coefficient, (ii) the image coefficient, (iii) the audio coefficient, and (iv) the video coefficient, and calculating a weighted number to describe the CoAC;

- (b) filtering or denying computer access to objects stored on computer-readable media that exceed the tolerated threshold for adult content.

The method can be used to classify objects stored on computer-readable media based on any combination of the above data type coefficients, and possibly other criteria.

Another object of the present invention is to use one or more of the classification methods outlined in detail below to classify Internet objects and objects stored on computer-readable media on the basis of any other identified criteria or parameter besides adult content.

Numerous additional features and aspects of the methods and device of the present invention are set forth in the detailed description that follows.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

The invention is described below in greater detail with reference to the accompanying drawings, which illustrate a preferred embodiment of the invention, and wherein:

Figure 1 is a block diagram of a computer system that can be used to implement one or more different embodiments of the present invention;

Figure 2 is a flow chart of the calculation of the eight coefficients and the calculation of the weighted sum or sums used to classify an Internet object;

Figure 3 is a flow chart of the calculation of the descriptor coefficient for a PICS code;

Figure 4 is a flow chart of the calculation of the name coefficient;

Figure 5 is a flow chart of the calculation of the text coefficient;

Figure 6 is a flow chart of the calculation of the image coefficient;

Figure 7 is a flow chart of the calculation of the audio coefficient;

Figure 8 is a flow chart of the calculation of the video coefficient;

Figure 9 is a flow chart of the calculation of the plug-in coefficient;

Figure 10 is a flow chart of the calculation of the relational coefficient;

Figure 11 is a flow chart of the steps involved in classifying and filtering access to an object containing adult content stored on computer-readable media; and

Figure 12 is a flow chart of the calculation of the four coefficients and the calculation of the weighted sum used to classify an object stored on computer-readable media.

## DESCRIPTION OF THE PREFERRED EMBODIMENTS

The device which performs the method of classification of Internet objects and objects stored on computer-readable media as described herein is illustrated in



Figure 1. Figure 1 is a high-level block diagram of a system (2) that can be used to implement one or more different embodiments of the present invention. As illustrated, the system includes a computer system (2) which itself comprises an input/output interface (4), a processor (6), and a memory (8) all conventionally interconnected by bus (10). Memory (8), which generally includes different modalities, all of which are not specifically shown for simplicity, may be implemented using random access memory (RAM), hard disk storage and/or other removable media devices, such as floppy disks, magnetic tape and CD-ROM drives, which are capable of reading information stored on a computer-readable medium such as a CD-ROM, CD-R, and CD-RW. In accordance with the present invention, the memory (8) stores an operating system (O/S) (12) and a variety of application programs (14). The application programs include an Internet browser (16), a search engine (18), and a classification module (20).

The computer system is coupled to a plurality of data sources via the input/output interface (4) and a bi-directional bus (10). The data sources include user input devices (22) such as a keyboard and/or mouse, a plurality of external databases (24), and a plurality of additional external information sources, including those accessed via the network (30). The additional information sources, of which only a few are specifically shown (24, 30), can include, e.g., an Internet connection via the network (30) and a broadcast receiver such as a television or satellite receiver. Inasmuch as the present invention will function with stored information,

e.g., data, regardless of its source, the particular modality through which such data is physically provided to the computer system is immaterial.

In addition to coupling the computer system to a plurality of external data sources, the input/output interface (4) electrically connects and interfaces the computer system to a plurality of output devices (26). The output devices could include, e.g., a display, such as a conventional colour monitor, and printer (28), such as a conventional laser printer or other well-known printer. In addition, the input/output interface (4) may be used to electrically connect and interface the computer system to a plurality of memory storage devices (8).

A user can invoke an application program or software module implementing the present invention through appropriate commands entered via user input devices (22) and/or in response to selecting a program icon appearing on display (26) that represents the application.

The first exemplary embodiment of the invention classifies Internet objects by using an automated web robot whose goal is to find Internet pages with adult content. In order to understand how the robot navigates the Internet, it is necessary to understand the differences in the levels of addressing required to identify Internet objects. At the highest level, Internet addresses are governed by unique "domain names", such as [playboy.com](http://www.playboy.com). Once granted this domain name, the owners of the domain name can set up computers that post Internet content at the "website" located at <http://www.playboy.com/>. This Internet content takes the form of

individual web pages. An Internet object that contains links to individual web pages is known as a directory. An example of a directory is <http://www.playboy.com/sex/>. Each individual web page has a unique address, known as a universal resource locator, or URL. An example of a URL is <http://www.playboy.com/sex/morsex.html>. The robot works by visiting and classifying Internet objects at individual URLs.

The robot takes as input a list of URLs and visits each one of them. At each site, the robot performs the classification analysis set out in detail below. At each site, the robot determines whether to add any new URLs found to its visitation list. This analysis is based on the level of adult content found on the URL. The robot maintains a list of URLs already visited so that no URL is visited more than once.

The robot builds and maintains a database of URLs which contain Internet objects which exceed the tolerated threshold for adult content. The database can be referenced upon executing Internet access via any communications port in any type of computer. Upon execution of Internet access, the database of URLs with adult content is referenced in order to block or filter access to Internet addresses that have been identified as having adult content that exceeds the tolerated threshold.

The invention uses a combination of text, image, video, sound, and relational analysis to determine whether a particular Internet object contains adult content, all without the need for human assessment. For the purposes of this invention, and all other embodiments of the invention and/or sub-inventions described below, adult

content is defined as anything not suitable for viewing by a minor, including violence/profanity, partial nudity, full nudity, sexual acts, gross depictions, intolerance, drugs/drug culture, militant/extremist propaganda, illegal acts, and alcohol/tobacco advertisements.

Figure 2 illustrates the calculation of the eight coefficients and the calculation of the weighted sum or sums used to classify an Internet object. The invention automatically computes, without human intervention, seven different values, or data type coefficients (audio coefficient (202), video coefficient (204), plug-in coefficient (206), descriptor coefficient (208), name coefficient (210), text coefficient (221), and image coefficient (214)). The steps for calculating the various data type coefficients are described in greater detail below.

The invention then combines the various data type coefficients, using a weighting scheme, into one value, (the first weighted sum (216)), called the coefficient of adult content ("CoAC") (216). The weighting scheme for the first weighted sum (216) can be determined automatically or can be adjusted by the user to reflect the relative acceptability of the adult content signalled by the various data type coefficients. The weighting scheme can also be adjusted to take account of other factors, such as the prevalence of various data types in an Internet object. For example, in a Internet page (defined as any logical unit of Internet content that can be downloaded using any transmission protocol) that contains 10 pictures and 5

words, more weight could be given to the image data type coefficients in the computed CoAC for the object.

After computing the first seven data type coefficients, and producing the first weighted sum (the CoAC (216)), the CoAC (216) is compared against a user-specified threshold value (218). Any object with a CoAC (216) exceeding the threshold value is classified as containing adult content, and is added (in step 220) to a database containing the URLs of sites which contain Internet objects that exceed the tolerated threshold for adult content.

Internet objects that are not classified as adult content on the basis of the CoAC (216) computed using the first seven data types (*i.e.* the audio coefficient (202), video coefficient (204), plugin coefficient (206), descriptor coefficient (208), name coefficient (210), text coefficient (221), and image coefficient (214)), are classified using a relational coefficient (222).

Once the relational coefficient (222) is computed, using the steps described in greater detail below, a second weighted sum (224) is computed, using the relational coefficient (222), as well as the other seven data type coefficients (*i.e.* the audio coefficient (202), video coefficient (204), plugin coefficient (206), descriptor coefficient (208), name coefficient (210), text coefficient (221), and image coefficient (214)). The weighting scheme for the second weighted sum (224) can be determined automatically or can be adjusted by the user to reflect the relative acceptability of the adult content signalled by the various data type coefficients. The weighting scheme

can also be adjusted to take account of other factors, such as the prevalence of various data types in an Internet object. For example, in a Internet page (defined as any logical unit of Internet content that can be downloaded using any transmission protocol) that contains 10 pictures and 5 words, more weight could be given to the image data type coefficients in the computed CoAC for the object.

After computing the eight data type coefficients, and producing the second weighted sum (the CoAC (224)), the CoAC (224) is compared against a user-specified threshold value (226). Any object with a CoAC (224) exceeding the threshold value is classified as containing adult content, and is added (in step 220) to a database containing the URLs of sites which contain Internet objects that exceed the tolerated threshold for adult content. Any object with a CoAC (224) below this threshold is ignored.

The method can be used to classify Internet objects based on any combination of the data type coefficients described below, and possibly other criteria.

### **The Descriptor Coefficient**

Many Internet objects have identifiers embedded in the object, known as meta data. Meta data is information stored within the Internet object. It is not normally visible to the viewer of an Internet object, but is machine-readable, and provides information about the Internet object that can be used by computer applications. Examples of meta data include the computer language in which the Internet object is

written, the author or creator of the Internet object, and keywords that describe the Internet object.

Internet content providers currently have the option of embedding voluntary rating labels in the meta data of their Internet objects. There exist several such labelling systems. All of these systems allow an analysis of the content provider's opinion on whether the object contains adult content. The emerging standard in labelling systems is the PICS label system, whereby providers of Internet content voluntarily rate their Internet objects in terms of the amount of adult content (namely nudity, sex, violence and profane language) contained in the Internet object. A complete description of the rating guide for PICS values can be found at: <http://www.icra.org/ratingsv01.html>. In essence, however, the nudity, sex, violence and profane language in the content of each Internet object is rated on a scale of zero to four, with four representing the most adult content in each of the categories.

For example, the PICS label for the web page <http://www.playboy.com/index.html> consists of the following single line of meta text:

```
"<meta http-equiv="pics-label" content='(pics-1.1
"http://www.rsac.org/ratingsv01.html" l gen true
comment "RSACi North America Server" for
http://www.playboy.com on "1996.05.04T06:51-0800" r
(n 4 s 3 v 0 l 4))'>"
```

The actual PICS rating is found in the very last portion of the PICS label, and it tells a reader that the content provider of the Playboy website has rated the level of nudity on the site as a four out of four , the level of sex on the site sex as a three out of four, the level of violence on the site as a zero out of four, and the level of profane language on the site language as a four out of four.

Figure 3 illustrates the calculation of the descriptor coefficient for a PICS code. As illustrated, the PICS code (302) of an Internet object is read by the device of the invention and the individual elements of the PICS code, known as labels, are parsed (304). The labels are weighted to reflect the relative acceptability of the adult content signalled by the various items rated by the PICS label. The weighting scheme can be determined automatically or determined by the user. The ratings of the labels are multiplied by their weighting (306) and an average PICS code rating is determined by adding the weighted values (308) and dividing by the number of PICS labels (310). The pseudocode for computing the descriptor coefficient is as follows:

1. PICScoef = 0
2. for each label in the PICS
  - 2.1 PICScoef =PICScoef + (label value \* weight)
  - 2.2 PICScoef =PICScoef / #labels
3. output PICScoef



## The Name Coefficient

Internet objects have identifiers, such as file names, directory names, domain names, URLs, etc. In most cases, these identifiers are chosen in some meaningful way that makes it possible to gain information about the Internet objects from their names.

A list of keywords that are likely to indicate adult content is developed. The keyword list can be developed in a plurality of ways, including an analysis of the content of Internet objects with adult content, or by incorporation of a user-identified list of keywords. Each keyword is given a weight, such that a higher weight is given to words that are more likely to indicate adult content. Examples of keywords might be slang terms for sexual organs, slang terms for sexual acts, profanities, and slang terms for sexual orientations. Proper names for sexual organs or other such clinical terms are either not included in the list or given a very low weight.

Figure 4 illustrates the calculation of the name coefficient. The robot obtains the text file identifiers of an Internet object (402), and analyzes each name. The robot computes all possible substrings contained in the name being analyzed (404). The robot accesses and uses the keyword list in calculating the name co-efficient, by checking for matches of words in the keyword list (406). The robot then calculates the name coefficient by adding the weights of all weighted keywords found in all substrings (408).

The pseudocode for computing the name coefficient is as follows.

1. set namecoeff = 0
2. for each substring, *s*, in the object's name
  - 2.1. if *s* is in the name keyword list
    - 2.1.1. namecoeff = namecoeff + the weight of *s*
3. output namecoeff

### The Text Coefficient

Many Internet objects, such as email messages, text documents and web pages, contain a text component. In such cases, the text can be analyzed to determine if it contains adult content.

As set out above, a list of keywords that are likely to indicate adult content is developed. In addition, a list of phrases likely to indicate adult content is developed in order to capture separate words that may not indicate adult content on their own but may do so in a phrase (*e.g.* "18 or over"). The phrases for the keyword list can be developed in a plurality of ways, including an analysis of the content of Internet objects with adult content, or by incorporation of a user-identified list of phrases. Each keyword and phrase is given a weight, such that a higher weight is given to words that are more likely to indicate adult content.

Figure 5 illustrates the calculation of the text coefficient. The robot retrieves the text contained in an Internet object (502). The robot eliminates, where necessary,

any formatting from the text of the Internet object to obtain an unformatted text (504). Next, the robot reduces each individual word in the unformatted text to a stem (506), using any known stemming algorithm, for example the Porter stemming algorithm [Porter, M.F., 1980, An Algorithm for Suffix Stripping, *Program* 14(3): 132-137]. The resulting word stems are then compared to the weighted keyword list (508). In addition, the text is searched for key phrases that exactly match key phrases in the weighted key phrase list (510). The first step in developing the numerator of the text coefficient is to add the weights of all weighted keywords found in all stemmed words and the weights of all weighted key phrases found in the original text (512). The second step in calculating the numerator of the text coefficient is to multiply the total of the weights of all stemmed words and all key phrases by a constant, which may be determined automatically or determined by the user (514). The constant, which is a mathematical function of the total number of words in the Internet object, functions as a scaling factor (or weighting factor). The denominator of the text coefficient is the number of words in the Internet object (516). The text coefficient is calculated by dividing the numerator by the denominator.

The pseudocode for this process is as follows.

1. set textcoeff = 0
2. remove all formatting from the text
3. for each word, w, in the text do

- 3.1. reduce  $w$  to a canonical form to get a word-stem,  $s$
- 3.2. if  $s$  is in the stem keyword list then
  - 3.2.1. set  $\text{textcoeff} = \text{textcoeff} + \text{the weight of } s$
4. for each phrase,  $p$ , in the key phrase list do
  - 4.1. let  $n$  = the number of occurrences of  $p$  in the text
  - 4.2. set  $\text{textcoeff} = \text{textcoeff} + n * \text{the weight of } p$
5. set  $\text{textcoeff} = \text{textcoeff} * k / \text{number of words in the text}$ , where  $k$  is a user-defined or automatically generated constant value
6. output  $\text{textcoeff}$

### The Image Coefficient

When an Internet object contains image data, or is an image itself, the image data can be analyzed to determine whether it contains adult content. Computation of the image coefficient proceeds using the following steps. First, a range of Hue-Saturation-Value ("HSV") colour space values that are classified as skin tones is developed, and provided to the robot for access and use during the classification process. [For an explanation of HSV colour space, see Gonzales and Woods, "Digital Image Processing", Addison Wesley, Reading MA, 1992.]

Skin tone ranges can be developed in a plurality of ways, including reliance on prior art. [See, for example, Forsyth and Fleck, "Automatic Detection of Human Nudes", *International Journal of Computer Vision* 32/1, pp. 63-77, 1999]. The invention can use a skin tone range in which skin tone is defined as any pixel with Hue values between 0 to 25 (of a maximum 180) and Saturation values between 50 and 230 (of a

maximum 255). This range may be further refined by statistical analysis of data settings, as optimal parameters depend on the data set being analyzed. The preferred embodiment of the invention uses a skin tone range in which skin tone is defined as any pixel with Hue values between 2 to 18 (of a maximum 180) and Saturation values between 80 and 230 (of a maximum 255). This range may also be further refined by statistical analysis of data settings, as optimal parameters depend on the data set being analyzed.

Figure 6 illustrates the calculation of the image coefficient. When the robot is reviewing images in an Internet object, the robot begins by converting the colour of the Internet object into the HSV colour space (602). The robot then compares every pixel of the Internet object image to identify which pixels fall within the skin tone ranges it has been provided (604).

The identification of such skin tone ranges is itself an improvement of the prior art, in that the transformation from Red/Green/Blue colour space to the HSV colour space before analysis allows a more accurate identification of skin type pixels.

The robot then divides the number of skin tone pixels in the Internet object image by the total number of pixels in the image (606). The robot compares the resulting value against an automatically generated or user-defined threshold value (608). If the resulting ratio is less than the threshold proportion of skin pixels, then the robot disregards the image (610). Each image in an Internet object that meets or exceeds the threshold proportion of skin pixels is assigned a nudity coefficient equal

to the percentage of skin pixels in the image (612). The nudity coefficient of each image in an Internet object can also be weighted (614) using a plurality of factors, including, but not limited to, the size of the image (with larger pictures getting a heavier weight) and the encoding type (with images in formats with more accurate colour ranges getting a higher weight).

The image coefficient is calculated by adding the weights of all weighted nudity coefficients generated by an Internet object (616), and then multiplying that total by a constant, which may be determined automatically or determined by the user (618). The constant, which is a mathematical function of the total number of images in the Internet object, functions as a scaling factor (or weighting factor).

The pseudocode for this process is as follows.

1. set imagecoeff = 0
2. for each image, i, in the object do
  - 2.1. set skincount = 0
  - 2.2. for each pixel, p, of i do
    - 2.2.1. if  $2 \leq \text{hue}(p) \leq 18$  and  $80 \leq \text{saturation}(p) \leq 230$  then
      - 2.2.1.1. set skincount = skincount + 1
  - 2.3. set skincount = skincount / number of pixels in i
  - 2.4. set imagecoeff = imagecoeff + skincount \* weight
3. if number of images > 0
  - 3.1. set imagecoeff = imagecoeff /  $\log_2$  (number of images)

#### 4. output imagecoeff

### **The Audio Coefficient**

Certain audio data can be transcribed into text, and the resulting transcription can be analyzed and classified, using the textual classification methods previously described. Figure 7 illustrates the calculation of the audio coefficient. To obtain the audio coefficient, audio components in the Internet object are converted to unformatted text (702) using any known speech recognition algorithms (for example Dragon Naturally Speaking available from Dragon Systems, Inc.). The resulting texts are then put together (concatenated) (704) and their text coefficient is computed in the same manner as the method discussed with respect to the text coefficient above (706). The result is the audio coefficient.

The pseudocode for this process is as follows.

1. set audiocoeff = 0
2. set text = null
3. for each audio stream, s, in the object do
  - 3.1. convert s to text to obtain text t
  - 3.2. append t to text
4. set audiocoeff = the text coefficient of extracted text
5. output audiocoeff

### **The Video Coefficient**

Video data contains components of both audio and image data. Figure 8 illustrates the calculation of the video coefficient. The invention analyzes video data

by first dividing the video data into image data (802) and audio data (804) as required by the media type. The resultant image data is evaluated using the image coefficient method (806), and an image coefficient is assigned for each frame of video data. The resultant audio data is evaluated using the audio coefficient method (808), and an audio coefficient is assigned for the audio data. The video coefficient is calculated by adding a weighted sum of the image and audio coefficients (810). The weighting can be done in a plurality of ways, including a weight based on the relative proportion of audio and image data found in each video object. In addition, the weighting scheme can be adjusted by the user to reflect the relative acceptability of the adult content signalled by the image and audio coefficients.

The pseudocode for this process is as follows.

1. set videocoeff = 0
2. let images = null
3. let audio = null
4. for each video component, v, in the object do
  - 4.1. add each frame of v to images
  - 4.2. add (concatenate) the audio stream of v to audio
5. set videocoeff = the image coefficient of images / (total number of frames)
6. set audiocoeff = the audio coefficient of the extracted audio data
7. set videocoeff = videoweight \* videocoeff + audioweight \* the audiocoeff of audio



## 8. output videocoeff

### The Plug-in Coefficient

As technology evolves, new types of Internet objects are introduced. Producers of Internet software, most notably Netscape and Microsoft, have responded to this trend by introducing plug-ins. Plug-ins are programs that operate with viewing software, in this case web browsers and mail readers, and allow users to view Internet object types for which viewers were not originally included in the software. In most cases, these plug-ins are provided by the organization that developed the Internet object type. Current examples include Adobe's Acrobat Reader and Macromedia's Flash Player.

The invention contemplates using information about new Internet object types by allowing a third party to produce plug-ins that can compute type specific adult content coefficients for new object types. The invention would obtain plug-in data from the Internet object (902). The invention would then weight the plug-in data (904). The weighting can be done in a plurality of ways, including a weight based on the popularity of the object type for use in association with Internet objects containing adult content. In addition, the weighting scheme can be adjusted by the user to reflect the relative acceptability of the adult content signalled by the various plug-ins. The plug-in coefficient would be calculated by adding the weighted plug-in data (906), and dividing by the number of plug-ins present (908).

The pseudocode for this process is as follows.

1. set plugincoeff = 0
2. for each plugin, p, appearing in the object do
  - 2.1. set plugincoeff = plugincoeff + adult content coefficient of p \* weight
3. set plugincoeff = plugincoeff / the number of plugins
4. output plugincoeff

### The Relational Coefficient

Although text, image, audio and video analysis will find the vast majority of sites containing adult content, some sites with adult content do not contain any of the preceding types of data. For example, some sites contain very extensive listings of pornography websites, but themselves contain little text, image, audio or video data that would allow the site to be discovered using the preceding methods.

Accordingly, for Internet objects that have not yet crossed the threshold of objectionable adult content, the robot performs one final analysis based on relations between Internet objects. There are two types of relations that the robot considers: "part-of" relations, and "links-to" relations. An Internet object "A" is said to be "part-of" another Internet object "B" if A is contained in B, in some natural sense. Examples include: files in a directory, directories on a machine, machines in an Internet domain, and messages in a newsgroup. Of course, "part-of" relationships are transitive. In other words, if A is a part of B and B is a part of C, then A is a part of C. An Internet object A is said to "link to" another Internet object B if A contains a reference to B. Examples include web pages that link to other web pages and email or newsgroup postings that are replies to previous postings.

Figure 10 illustrates the calculation of the relational coefficient. First, the robot compiles a list of all of the other Internet objects that an Internet object under consideration links to (the “linked-to” objects) (1002), as well as a list of all of the other Internet objects that an Internet object under consideration is part of (the “part of” objects) (1004).

WHITENER\4026861\1

in the list of objects that exceed the tolerated threshold for adult content is calculated (1012), and a similar calculation is performed for "part of" objects (1016).

The relational coefficient is computed by calculating a weighted average of the average CoAC of the Internet objects in the lists of Internet objects that exceed the tolerated threshold for adult content (1020). The weighting scheme, which can be determined automatically or set by the user, can take a plurality of forms. For example, the weighting scheme may be adjusted by the user to reflect the relative acceptability of the adult content signalled by the various data type coefficients which make up the CoAC of the "linked to" objects and the "part of" objects.

This process is given by the following pseudocode.

1. set linktocoef = 0
2. for each object, o, having a "links-to" relationship with the object do
  - 2.1. set linktocoef = linktocoef + the CoAC of o
3. set linktocoef = linktocoef / # objects having a "links-to" relationship with the object
4. set containscoef = 0
5. for each object, o, having a "contains" relationship with the object do
  - 5.1. set containscoef = containscoef + the CoAC of o
6. set containscoef = containscoef / # objects having "contains" relationship with the object
7. set relcoef = (linkweight \* linktocoef + containsweight \* containscoef)/2
8. output relcoef

## Gathering of Data

The analysis set out above has been implemented and incorporated into a web robot whose goal is to find web pages with adult-content. In this section, we describe the implementation of this robot. The robot takes as input a list of URLs and visits each one of them in an attempt to determine whether they are adult content sites or not. The initial URLs are sites with a great number of links to known adult content sites. The robot uses a queue,  $Q$ , of URLs to visit, and operates as follows:

1. initialize  $Q$  with input list of URLs
2. while  $Q$  is not empty do
3. choose a random URL,  $u$ , from  $Q$
4.   if  $u$  appears to be an adult-content site then
5.   add any links to HTML pages found on  $u$  to  $Q$
6.   end if
7. end while

The robot maintains a list of URLs already visited so that no URL is visited more than once. Line 4 of the above algorithm is implemented as follows [where the site (hostname) of URL is denoted by site (URL); the directory of the URL on the host is denoted by directory (URL); and the domain is denoted as domain (URL) For example, if URL were `http://www.scs.carleton.ca/~morin/friends.html`, then site (URL) would be `http://www.scs.carleton.ca`, directory (URL) would be `http://www.scs.carleton.ca/~morin/`, and domain would be `carleton.ca`.]:

1. if the web page domain(*URL*) appears to be adult content then
2. Mark *URL* as adult content.
3. if the web page *site(URL)* appears to be adult content then
4. mark *URL* as adult-content
5. else if the web page *directory(URL)*  
appears to be adult content then
6. mark *URL* as adult content
7. else if the web page *URL* appears to  
be adult-content then
8. mark *URL* as adult content
9. else mark *URL* as not adult content

If a site or directory becomes classified as adult-content then all URLs on that site or in that directory will also be classified as adult content, but the converse is not true. The robot maintains a database of URLs of sites that contain Internet objects that exceed the tolerated threshold for adult content. The database can be referenced upon executing Internet access via any communications port in any type of computer. Upon execution of Internet access, the database of URLs with adult content is referenced in order to block access to Internet addresses that have been identified as having adult content that exceeds the tolerated threshold.

It will be understood by those skilled in the art that the invention has uses in generating adult content lists for Internet filtering software. As well, the invention has application in finding and filtering Internet content as it passes through the TCP/IP stack. Sub inventions have applications in automatic censoring of pictures,

movies, television, radio and other media. By sub inventions, we mean each novel data type analysis, and each novel combination of data type analyses disclosed herein. In addition, the invention and/or sub-inventions have applications in the identification and classification of Internet objects on the basis of criteria and parameters other than adult content. Accordingly, it is to be understood that the present invention is not limited to the various embodiments set forth above for illustrative purposes, but also encompasses various modifications which are rendered obvious by the present disclosure.

The second exemplary embodiment of the invention classifies objects stored on computer-readable media by using a classification module resident on a computer system to review and classify objects stored on computer-readable media, either in a pre-processing phase or upon access to an object. The invention uses a combination of text, image, video, and audio analysis to determine whether a particular object stored on computer-readable media contains adult content, all without the need for human assessment. Figure 11 illustrates the steps involved in this process. Upon receipt of an instruction from a user to access an object (1102), the operating system would access the media (1104), and the classification module would perform the classification analysis (1106). If an object exceeded the tolerated threshold for adult content, the classification module would filter or deny access to the object (1108). Of course, if the object did not exceed the tolerated threshold for adult content, the classification module would allow access to the object (1110).

00000000-0000-0000-0000-00000000

Figure 12 is a more detailed diagram of the calculation of the four coefficients and the calculation of the weighted sum used to classify an object stored on computer-readable media. The invention automatically computes, without human intervention, four different values, or data type coefficients (audio coefficient (1202), video coefficient (1204), text coefficient (1206), and image coefficient (1208)). The steps for calculating the various data type coefficients are described below.

The invention then combined the various data type coefficients, using a weighting scheme, into one value (the weighted sum (1210), called the coefficient of adult content (CoAC). The weighting scheme for the weighted sum can be determined automatically or can be adjusted by the user to reflect the relative acceptability of the adult content signalled by the various data type coefficients. The weighting scheme can also be adjusted to take account of other factors, such as the prevalence of various data types in an object stored on computer-readable media. For example, in an object that contains 10 pictures and 5 words, more weight could be given to the image data type coefficients in the computed CoAC for the object.

After computing the four data type coefficients, and producing the weighted sum (CoAC (1210)), the CoAC (1210) is compared against a user-specified threshold value (1212). Any object with a CoAC (1210) exceeding this threshold is classified as containing adult content, and the classification module would filter or deny access to the object (1214). Of course, if the object did not exceed the tolerated threshold for adult content, the classification module would allow access to the object (1216).



The classification module (20) will analyze the data being requested. The classification module (20) will determine, based on user-definable criteria, to either allow or disallow access to the storage devices (32) for read or write access. For example, read access to the CD-ROM may be denied because the content on the CD-ROM contains objects classified as adult content. Similarly, the classification module could be invoked when an attempt is made to write to a storage device. The classification module could then allow or disallow write access to the storage device depending on the classification of the data to be written. In an identical manner, control to any item via the input/output interface (4) is possible using the classification module (20) and the operating system (12). For example, access to a printer (28) or network device (30) can be controlled by the classification module (20) in place of the storage devices (32) in the above example.

The method can be used to classify objects stored on computer-readable media based on any combination of the above data type coefficients, and possibly other criteria.

### The Text Coefficient

Many objects stored on computer-readable media, such as text documents, contain a text component. In such cases, the text can be analyzed to determine if it contains adult content.

As set out above with respect to the calculation of the text coefficient in the classification of Internet objects, a list of keywords and phrases that are likely to indicate adult content is developed. The classification module is given access to the keyword list, which is stored on a database resident on the computer system. Figure 5 illustrates the calculation of the text coefficient. When the classification module reviews an object stored on computer-readable media (502), it begins its analysis by eliminating, where necessary, any formatting from the text of the object to obtain an unformatted text (504). Next, the classification module reduces each individual word in the unformatted text to a stem (506), using any known stemming algorithm, for example, the Porter stemming algorithm [Porter, M.F., 1980, An Algorithm for Suffix Stripping, *Program* 14(3): 132-137]. The resulting word stems are then compared to the weighted keyword list (508). In addition, the text is searched for key phrases that exactly match key phrases in the weighted key phrase list (510). The first step in developing the numerator of the text coefficient is to add the

weights of all weighted keywords found in all stemmed words and the weights of all weighted key phrases found in the original text (512). The second step in calculating the numerator of the text coefficient is to multiply the total of the weights of all stemmed words and all key phrases by a constant, which may be determined automatically or determined by the user (514). The constant, which is a mathematical function of the total number of words in the Internet object, functions as a scaling factor (or weighting factor). The denominator of the text coefficient is the number of words in the Internet object (516). The text coefficient is calculated by dividing the numerator by the denominator.

The pseudocode for this process is as follows.

1. set textcoeff = 0
2. remove all formatting from the text
3. for each word, w, in the text do
  - 3.1. reduce w to a canonical form to get a word-stem, s
  - 3.2. if s is in the stem keyword list then
    - 3.2.1. set textcoeff = textcoeff + the weight of s
4. for each phrase, p, in the key phrase list do
  - 4.1. let n = the number of occurrences of p in the text
  - 4.2. set textcoeff = textcoeff + n \* the weight of p
5. set textcoeff = textcoeff \* k / number of words in the text, where k is a user-defined or automatically generated constant value
6. output textcoeff

## The Image Coefficient

When an object stored on computer-readable media contains image data, or is an image itself, the image data can be analyzed to determine whether it contains adult content. Computation of the image coefficient proceeds using the steps set out in the discussion of the image coefficient with respect to classification of Internet objects.

Figure 6 illustrates the calculation of the image coefficient. When the classification module is reviewing images in an object stored on computer-readable media, the module begins by converting the colour of the object into the HSV colour space (602). The classification module then compares every pixel of the object image to identify which pixels fall within the skin tone ranges it has been provided (604). The module then divides the number of skin tone pixels in the object image by the total number of pixels in the image (606). The module compares the resulting value against an automatically generated or user-defined threshold value (608). If the resulting ratio is less than the threshold proportion of skin pixels, then the classification module disregards the image (610). Each image in an object stored on computer-readable media that meets or exceeds the threshold proportion of skin pixels is assigned a nudity coefficient equal to the percentage of skin pixels in the image (612). The nudity coefficient of each image in an object stored on computer-readable media can be weighted (614) using a plurality of factors, including, but not limited to, the size of the image (with larger pictures getting a heavier weight) and

the encoding type (with images in formats with more accurate colour ranges getting a higher weight).

The image coefficient is calculated by adding the weights of all weighted nudity coefficients generated by an object stored on computer-readable media (616), and then multiplying that total by a constant, which may be determined automatically or determined by the user (618). The constant, which is a mathematical function of the total number of images in the object, functions as a scaling factor (or weighting factor).

The pseudocode for this process is as follows.

1. set imagecoeff = 0
2. for each image, i, in the object do
  - 2.1. set skincount = 0
  - 2.2. for each pixel, p, of i do
    - 2.2.1. if  $2 \leq \text{hue}(p) \leq 18$  and  $80 \leq \text{saturation}(p) \leq 230$  then
      - 2.2.1.1. set skincount = skincount + 1
  - 2.3. set skincount = skincount / number of pixels in i
  - 2.4. set imagecoeff = imagecoeff + skincount
3. if number of images > 0
  - 3.1. set imagecoeff = imagecoeff /  $\log_2$  (number of images)
4. output imagecoeff

### The Audio Coefficient

As previously stated, certain audio data can be transcribed into text, and the resulting transcription can be analyzed using the textual classification method previously described. Figure 7 illustrates the calculation of the audio coefficient. To obtain the audio coefficient, audio components in the object stored on computer-readable media are converted to unformatted text (702) using any known speech recognition algorithms (for example Dragon Naturally Speaking available from Dragon Systems, Inc.). The resulting texts are then put together (concatenated) (704) and their text coefficient is computed in the same manner as the method discussed with respect to the text coefficient above (706). The result is the audio coefficient.

The pseudocode for this process is as follows.

1. set audiocoeff = 0
2. set text = null
3. for each audio stream, s, in the object do
  - 3.1. convert s to text to obtain text t
  - 3.2. append t to text
4. set audiocoeff = the text coefficient of extracted text
5. output audiocoeff

### The Video Coefficient

Video data contains components of both audio and image data. Figure 8 illustrates the calculation of the video coefficient. The invention analyzes video data by first dividing the video data into image data (802) and audio data (804) as

required by the media type. The resultant image data is evaluated using the image coefficient method (806), and an image coefficient is assigned for each frame of video data. The resultant audio data is evaluated using the audio coefficient method (808), and an audio coefficient is assigned for the audio data. The video coefficient is calculated by adding a weighted sum of the image and audio coefficients (810). The weighting can be done in a plurality of ways, including a weight based on the relative proportion of audio and image data found in each video object. In addition, the weighting scheme can be adjusted by the user to reflect the relative acceptability of the adult content signalled the image and audio coefficients.

The pseudocode for this process is as follows.

1. set videocoeff = 0
2. let images = null
3. let audio = null
4. for each video component, v, in the object do
  - 4.1. add each frame of v to images
  - 4.2. add (concatenate) the audio stream of v to audio
5. set videocoeff = the image coefficient of images / (total number of frames)
6. set audiocoeff = the audio coefficient of the extracted audio data
7. set videocoeff = videoweight \* videocoeff + audioweight \* the audiocoeff of audio
8. output videocoeff

## Gathering of Data

The analysis set out above has been implemented and incorporated into a classification module that reviews and classifies objects stored on computer-readable media, with the goal of filtering or blocking access to objects that exceed a user-defined tolerance for adult content.

It will be understood by those skilled in the art that the invention has uses in filtering or blocking access to objects stored on computer-readable media that exceed user-defined thresholds for adult content. Sub inventions have applications in automatic censoring of pictures, movies, television, radio and other media. By sub inventions, we mean each novel data type analysis and each novel combination of the data type analyses disclosed herein. In addition, the invention and/or sub-inventions have applications in the identification and classification of objects stored on computer-readable media on the basis of criteria and parameters other than adult content. Accordingly, it is to be understood that the present invention is not limited to the various embodiments set forth above for illustrative purposes, but also encompasses various modifications which are rendered obvious by the present disclosure.

00000000-0000-0000-0000-00000000